Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence (IJCAI 2015)

Weakly Supervised RBM for Semantic Segmentation

Yong Li, Jing Liu, Yuhang Wang, Hanqing Lu, Songde Ma
National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences
{yong.li,jliu,yuhang.wang,luhq}@nlpr.ia.ac.cn, masd@most.cn

Abstract

In this paper, we propose a weakly supervised Restricted Boltzmann Machines (WRBM) approach to deal with the task of semantic segmentation with only image-level labels available. In WRBM, its hidden nodes are divided into multiple blocks, and each block corresponds to a specific label. Accordingly, semantic segmentation can be directly modeled by learning the mapping from visible layer to the hidden layer of WRBM. Specifically, based on the standard RBM, we import another two terms to make full use of image-level labels and alleviate the effect of noisy labels. First, we expect the hidden response of each superpixel is suppressed on the labels outside its parent image-level label set, and a non-image-level label suppression term is formulated to implicitly import the image-level labels as weak supervision. Second, semantic graph propagation is employed to exploit the cooccurrence between visually similar regions and labels. Besides, we deal with the problems of label imbalance and diverse backgrounds by adapting the block size to the label frequency and appending hidden response blocks corresponding to backgrounds respectively. Extensive experiments on two real-world datasets demonstrate the good performance of our approach compared with some state-of-the-art methods.

1 Introduction

Semantic image segmentation is a fundamentally challenging problem, aiming at assigning semantic labels to image regions [Xie *et al.*, 2014]. Compared with traditional image segmentation, it provides higher-level understanding about image contents. While compared with typical image classification, it provides more fine-grained semantic understanding of images. Therefore, image semantic segmentation provides a virtual solution to bridge the semantic gap, and becomes one of the core problems in computer vision [Shotton *et al.*, 2009][Zhang *et al.*, 2012].

In the past years, semantic image segmentation has attracted a lot of attention, and significant progress has been achieved [Shotton *et al.*, 2006; 2008; Liu *et al.*, 2009a; Vezhnevets *et al.*, 2012; Farabet *et al.*, 2013]. Although these methods have shown promising results, they rely on a training set of images with pixel-level labels. However, the high cost on the ground truth acquisition restricts the wide usage of such work. Fortunately, with the rapid spread of online photo sharing websites (e.g., Flicker), large numbers of images with image-level labels become available. These image-level labels can be further exploited to make semantic segmentation [Liu *et al.*, 2009b; Vezhnevets *et al.*, 2011; Xie *et al.*, 2014]. In contrast to fully supervised setting with pixel-level labels, it is more challenging to make weakly supervised semantic segmentation with image-level labels.

Recently, a few methods have been proposed to address the weakly supervised segmentation problem [Liu *et al.*, 2009b; 2013; Zhang *et al.*, 2013; Xie *et al.*, 2014]. In general, the core task of weakly supervised semantic segmentation is to learn the mapping between image label and low-level feature of local regions. [Liu *et al.*, 2009b] attempted to capture the cooccurrence by a bi-layer sparse coding model. [Zhang *et al.*, 2013] developed a way to learn the mapping by evaluating the classifier, in which a good classifier will have good reconstruction basis for positive samples and large reconstruction error for the negative samples. Meanwhile, Graph propagation based methods are proposed to restrict visually similar regions to have similar labels [Liu *et al.*, 2012; 2013; Xie *et al.*, 2014]. Generally, the existing methods initialize the label of superpixels with image-level labels explicitly and refine the model consequently to get the final label for each superpixel. Such an explicit importation of image-level labels as supervision may be crude a little, since the limited discriminative ability of superpixels tends to make the performance of label refinement unsatisfied to some extent. Thus, how to import the image-level labels as weak supervision is a challenging but valuable problem for semantic segmentation.

In this paper, we propose a Weakly supervised Restricted Boltzmann Machines (WRBM) to deal with the problem of semantic segmentation by exploring the image-level labels as shown in Figure 1. The images are first over segmented into superpixels, and the extracted feature of each superpixel is set as the input to the visible nodes of WRBM. The hidden nodes of WRBM are divided into several blocks, where each block corresponds to a specific label. Under such setting, the superpixel label can be decided by the hidden-node block with the maximum response. The proposed WRBM is an extended version of the standard RBM by introducing two

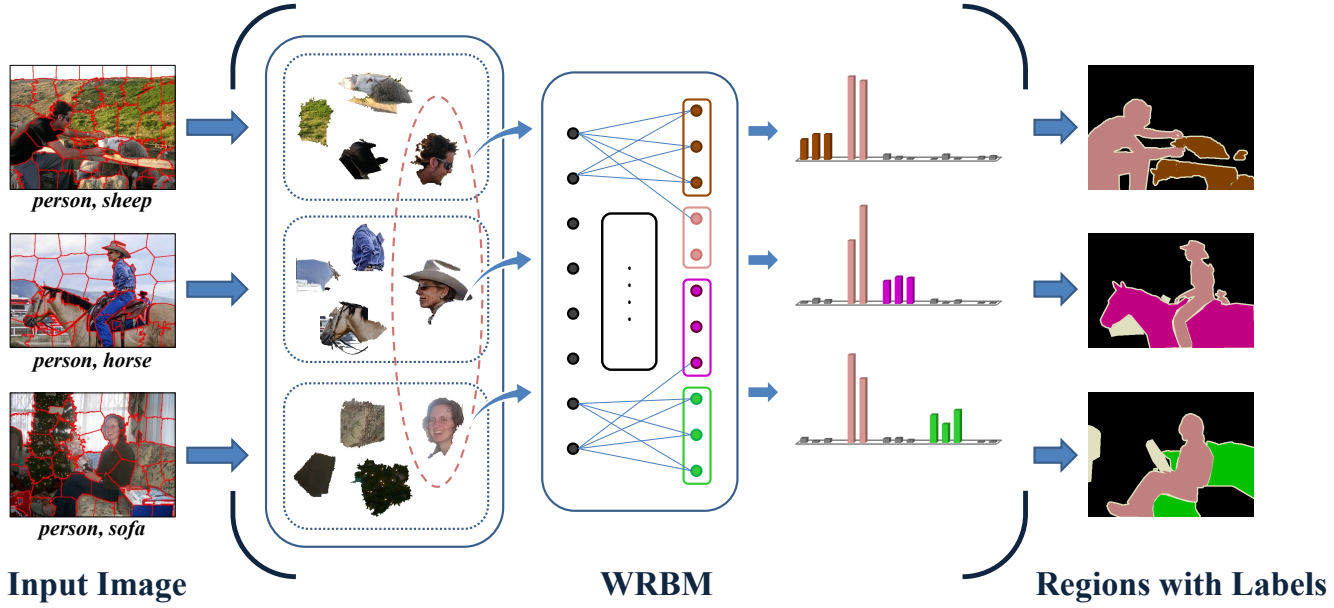

Figure 1: The overview of our approach. (Best viewed in color). (I) Oversegment each image into superpixels and extract features for each superpixel; (II) Learn the mapping associations between image labels and local regions by exploiting their cooccurrence among the training set via non-image-level label suppression. The hidden layer of WRBM is divided into multiple blocks and each block corresponds to a specific label; (III) Infer the label for local region with the learnt mapping associations.

additional terms in its objective function. The first term is a non-image-level label suppression one to implicitly import the image-level labels as weak supervision. Such implicit importation works on the intuition that the label of a superpixel is impossible to be the ones outside its parent image-level labels. Thus, we employ the term to suppress the response of blocks corresponding to those impossible labels. The second term is a semantic graph propagation one to make sure that visually similar superpixels sharing common image-level label have similar hidden response. Besides, we modify the model to deal with the problems of diverse backgrounds and label imbalance in the training dataset. For the former, we add multiple background blocks to the hidden layer, and assume each image corresponds to one background block. To deal with the label imbalance, we design the size of the block corresponding to each label to be an inversely correlated number with the label frequency. That is, we expect to amplify the response of an infrequent label by designing a larger block size. Finally, extensive experiments on two real-world datasets, i.e., PASCAL [Everingham *et al.*, 2010] and LabelMe [Russell *et al.*, 2008], demonstrate the effectiveness of our approach compared with some state-of-the-art methods. Generally, the main contributions are summarized as follows.

- We propose a RBM-based learning framework for the task of semantic segmentation. The correspondence between labels and hidden responses of WRBM gives a direct solution to the prediction of superpixel label.
- The non-image-level label suppression and the semantic graph propagation are employed together to make full use of the image-level labels and alleviate the effect of the noisy labels.
- The changeable block size and block number of the hidden layer are designed to handle the problems of label imbalance and diverse backgrounds of training data.

2 Related Work

Semantic segmentation has attracted wide interests due to its importance in bridging high-level concepts to low-level features of local regions. Most semantic segmentation approaches suppose that a training dataset with pixel-level labels is given [Vezhnevets *et al.*, 2012] [Shotton *et al.*, 2006] [Farabet *et al.*, 2013] [Liu *et al.*, 2009a] [Farabet *et al.*, 2013]. A typical way to model the problem with pixel-level label is based on Conditional Random Field (CRF) [Shotton *et al.*, 2006], The basic formulation is defined on pixel values with various potential functions, including shape, texture, color, location and edge cues. Lots of extensions are proposed to modify the CRF with high-order potentials [Ladicky *et al.*, 2009], hierarchical features [Farabet *et al.*, 2013], label cooccurrence [Ladicky *et al.*, 2010]. Another direction is to develop non-parametric methods to transfer labels from training images to the query image [Liu *et al.*, 2009a] [Liu *et al.*, 2011]. [Myeong and Lee, 2013] further explores high-order semantic relation with label transfer. All of these approaches require pixel-level labels for training, which are expensive to obtain in practise.

Weakly-supervised methods have emerged and attracted more attention due to the weak requirement of supervision [Vezhnevets *et al.*, 2011; 2012; Vezhnevets and Buhmann, 2010; Liu *et al.*, 2013; 2012; Zhang *et al.*, 2013; 2014;

Xie *et al.*, 2014]. [Vezhnevets and Buhmann, 2010] cast the semantic segmentation task as a multiple instance learning problem. They adopted semantic texton forest as the basic framework and extended it for the MIL setting. [Vezhnevets *et al.*, 2011] extended [Vezhnevets and Buhmann, 2010] with a multi-image model, in which smoothness between adjacent and similar superpixels is encouraged. [Vezhnevets *et al.*, 2012] exploited multiple visual cues in this weakly supervised setting with a parametric family of structured models. Meanwhile, [Liu *et al.*, 2009b] proposed a bi-layer sparse coding method, in which an image region is sparsely constructed with the regions of the same image-level label. What is more, [Liu *et al.*, 2012; 2013] developed a weakly supervised graph propagation model by considering superpixel consistency and weak supervision information simultaneously. [Xie *et al.*, 2014] further verified the importance of semantic graph construction in the graph propagation model.

Generally, the existing weakly supervised methods leverage the image-level labels explicitly, they initialize the label of superpixel with image-level labels and refine the model consequently to get the final label for each superpixel. Such methods are easily affected by the noise labels, which usually exist in real applications. Furthermore, different label concepts occur quite differently in the real case. Treating different labels equally without considering label imbalance will limit the performance of segmentation methods. Leveraging the weak image-level labels properly and addressing the problems above become the focus of this paper.

3 The Proposed Approach

We propose a weakly supervised RBM based method for semantic segmentation via non-image-level label suppression. RBM is an undirected graphical model, which consists of a visible layer and a hidden layer. It can model the data in a transformed subspace with the unsupervised setting. To address the semantic segmentation problem, units of the hidden layer are divided into multiple blocks, where each block corresponds to a specific label concept. Mapping associations between low-level features and hidden label blocks can be achieved via non-image-level label suppression and graph propagation.

Before the detailed discussion of each component, we first summarize some notations. Given a set of images τ, each image i is oversegmented into N_i superpixels [Achanta *et al.*, 2012]. Feature is extracted on each superpixel with the bag-of-words model using SIFT [Lowe, 2004] descriptor and color feature. x_{ij} denotes the feature of the j-th superpixel in image i. In addition, the image-level label set for the i-th image is denoted as S_i.

3.1 Non-image-level Label Suppression

Generally, the semantic segmentation problem is to learn the mapping from low-level features of local regions to high-level concepts. Unlike the fully supervised setting [Shotton *et al.*, 2006] [Farabet *et al.*, 2013], the problem under weakly supervised setting becomes very challenging because of the absence of pixel-level labels. For a specific image i, any label among the image-level label set S_i may be the truth for the superpixels. We can not learn a direct mapping from low-level features to high-level labels under the weakly supervised setting. However, the image-level label provides the important cue that there will be no mapping from the superpixels to the none-image-level image labels (labels not in S_i). To incorporate such idea into the standard RBM framework, we import a regularization term to suppress blocks corresponding to the none-image-level labels.

Standard RBM has a single layer of hidden units that are not connected to each other and have undirected, symmetrical connections to a layer of visible units [Hinton and Salakhutdinov, 2006; Bengio, 2009]. A joint configuration of the visible units **v** and hidden units **h** has an energy function as follows,

$$E_r(\mathbf{v}, \mathbf{h}) = -\mathbf{h}^T W \mathbf{v} - \mathbf{b}^T \mathbf{v} - \mathbf{c}^T \mathbf{h} \tag{1}$$

where W is the weight matrix between visible units and hidden units, **b** and **c** are the offsets of visible units **v** and hidden units **h**. The joint probability distribution of all the units is defined as follows,

$$P(\mathbf{v}, \mathbf{h}) = \frac{1}{Z} exp(-E_r(\mathbf{v}, \mathbf{h}))) \tag{2}$$

where Z is the partition function.

Given visible units **v**, the probability for the binary hidden unit $\mathbf{h}_m$ to be 1 can be obtained as follows,

$$P(\mathbf{h}_m = 1|\mathbf{v}) = \frac{1}{1 + \exp(\mathbf{c}_m + W_{m,:}\mathbf{v})} \tag{3}$$

where $W_{m,:}$ is the m-th row of the matrix W. Similarly, given hidden units **h**, the probability for the binary visible unit $\mathbf{v}_n$ to be 1 is as follows,

$$P(\mathbf{v}_n = 1|\mathbf{h}) = \frac{1}{1 + \exp(\mathbf{b}_n + W'_{:,n}\mathbf{h})} \tag{4}$$

where $W_{:,n}$ is the n-th column of the matrix W.

To import semantic information, the hidden layer units of RBM are divided into multiple blocks, and block B_k corresponds to the k-th label concept L_k. The response function for the units in block B_k is defined as follows,

$$E_{B_k} = \sum_{m \in B_k} \mathbf{h}_m^2 \tag{5}$$

where m is the index for the hidden unit in block B_k. For a specific image i with image-level label set S_i, response of the blocks corresponding to the non-image-level labels should be small. We capture this property by importing non-image-level suppression term as follows,

$$E_s = \sum_{i \in \tau} \sum_{j \in N_i} E_{B_{k \notin S_i}}. \tag{6}$$

As a result, mapping to the image-level labels will be encouraged for each superpixel with the suppression term.

Adaptive Block Size

The hidden units of RBM are divided into multiple blocks according to the label concept. A natural division method for the blocks is of the equal size. However, frequencies of different labels are usually highly imbalanced as shown in Figure 2. The background label 'sky' occurs in 85% images

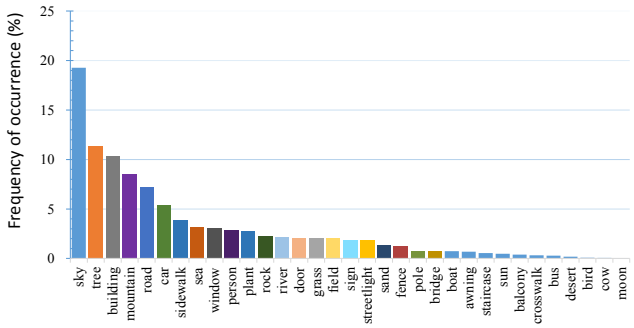

Figure 2: Label frequency on the LabelMe LMO dataset.

of the dataset, while the specific object 'moon' occurs in no more than 1% images. If the block is of the same size, images of high frequency labels will dominate the non-image-level suppression term of Equation (6). To deal with the label imbalance and emphasize the effect of low frequency labels, we increase the block size of the labels with low frequency directly. Empirically, the block size Q_{B_k} is modeled with respect to the label frequency as follows,

$$Q_{B_k} = M * \exp(-\frac{f_k}{\delta}) + C \tag{7}$$

where M is a hyper-parameter to control the scale of the block size, C is a smooth constant, δ is a free parameter to control the decay rate. f_k is the label frequency calculated from the training dataset as follows, $f_k = \frac{n_k}{\sum_{k'} n_{k'}}$, and n_k is the number of images containing label k.

Diverse Backgrounds

To deal with the diverse backgrounds, a intuitive idea is to add a block in the hidden layer corresponding to the class "background". However, the added block will have maximum response to the most commonly occurred visually similar superpixels among the dataset instead of the background regions for each image. Therefore, multiple blocks are preferred to deal with the diverse backgrounds. The background for a specific image belongs to one possible background block. A random background label is added to each image. With n_m multiple background blocks, the probability of n_i different images to have the same background label is $(\frac{1}{n_m})^{n_i-1}$, which is usually very small.

3.2 Semantic Graph Propagation

Label propagation plays an important role for weakly supervised semantic segmentation [Liu *et al.*, 2012; Zhang *et al.*, 2014; Xie *et al.*, 2014]. Semantic graph propagation term is proposed to make sure that similar superpixels sharing common image-level label have similar hidden response. If two similar local regions from different images share common label, then it is natural to tag these regions with the common label. [Xie *et al.*, 2014] validates the importance of affinity graph with kinds of construction methods. Some labels usually occur together, like "grass" and "sheep". It is difficult to label superpixels in such images. In order to embed more discriminative information into the affinity graph, we propose

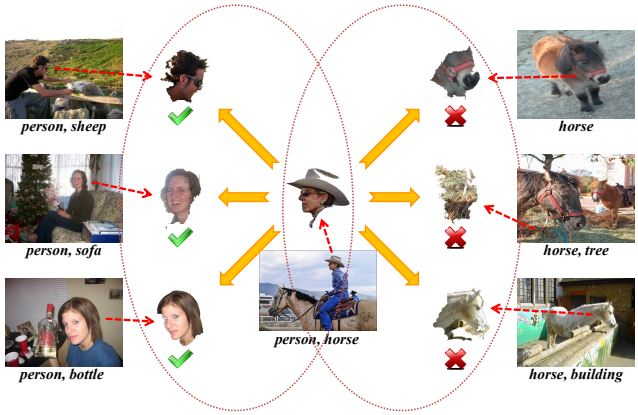

Figure 3: Illustration of semantic graph construction. For a specific superpixel in the image with labels "person" and "horse", the K nearest neighbors are found in the images with labels "person" and "horse" separately. Meanwhile the nearest images with the label "person" should not contain the label "horse", and the nearest images with the label "horse" should not contain the label "person". Finally, the KNN superpixels is selected with Equation 8.

a affinity graph construction method with semantic exclusion property. For a specific image i with label set S_i, a label L_i in S_i is denoted as $L_i \in S_i$, and its complementary set is denoted as L_i^{cs} with respect to S_i. For the superpixels in image i, we seek to find the K nearest neighbors for each label L_i separately. The K-NN superpixels are selected in the images, which contain label L_i, but do not contain the labels in L_i^{cs}. Figure 3 illustrates such process intuitively. Finally, the K-NN superpixels for a superpixel j in image i is decided with maximum semantic similarity,

$$\max_{L_i} \sum_{l \in K(ij)} A_{ij,l}^{L_i} \tag{8}$$

where $A_{ij,l}^{L_i} = \exp(-\frac{\|x_{ij}-x_l\|^2}{t})$ is the similarity measure for the superpixel pair with label L_i, and t is a free parameter to control the decay rate. The advantage of such semantic graph provides more discriminative information about the true label for each superpixel.

With the semantic affinity graph, the graph propagation term can be formulated as follows,

$$E_g = \sum_{i \in \tau} \sum_{j \in N_i} \sum_{l \in K(ij)} A_{ij,l} \|\mathbf{h}(x_{ij}) - \mathbf{h}(x_l)\|^2 \tag{9}$$

where x_{ij} is the extracted feature of the j-th superpixel in image i, and $K(ij)$ denotes the index set of the K nearest neighbors. $\mathbf{h}(x)$ is the hidden response vector corresponding to the visible input x.

3.3 Objective Function of WRBM

By incorporating the non-image-level label suppression term and semantic graph propagation term into the basic RBM framework, we can get the final energy function as follows,

$$E = E_r + \alpha E_s + \beta E_g \tag{10}$$

where α and β are the tradeoff parameters of the proposed two terms.

3.4 Semantic Segmentation with WRBM

Given that the proposed model WRBM is well learnt, the semantic segmentation process can be performed by finding the block with maximum response. Specifically, for any given image i with label set S_i, The pixel-level label T_{ij} of the superpixel x_{ij} can be assigned by maximizing the block response as follows,

$$T_{ij} = \max_{k \in S_i} E_{B_k} \tag{11}$$

where E_{B_k} is the response of block k given input feature x_{ij}, which is defined in Equation (5).

4 Optimization

For the basic problem of RBM, parameters can be estimated by minimizing the negative log-likelihood $-\sum_{\mathbf{h}} \log P(\mathbf{v}, \mathbf{h})$ via Contrastive Divergence [Hinton, 2002]. Contrastive Divergence is an approximation of the log-likelihood gradient that has been found to be a successful update rule for training RBM [Bengio, 2009]. Optimization algorithm of the proposed WRBM can be achieved by modifying Contrastive Divergence directly, since the non-image-level suppression term and graph propagation term are both convex and differentiable. The Contrastive Divergence algorithm with Weak Supervision (CDWS) is shown in Algorithm 1, where $\odot$ denotes element-wise product of two vectors.

Algorithm 1 CDWS

Input: Extracted feature $\mathbf{x}_{ij}$, image-level label set S_i, parameter weights α and β, learning rate ε, maximum epoch number E_{ch}, similarity matrix $A_{ij,l}$

Initialize: W, $\mathbf{b}$, $\mathbf{c}$
While $ite < E_{ch}$ **do**
$P(\mathbf{h}^1 = 1|\mathbf{x}_{ij}) = \frac{1}{1+\exp(\mathbf{c}+W\mathbf{x}_{ij})}$
sample $\mathbf{h}^1 \in \{0, 1\}$ from $P(\mathbf{h}|\mathbf{v})$
$P(\mathbf{x}_{ij} = 1|\mathbf{h}) = \frac{1}{1+\exp(\mathbf{b}+W^T\mathbf{h}^1)}$
sample $\mathbf{x}^2_{ij} \in 0, 1$ from $P(\mathbf{x}^2_{ij}|\mathbf{h}^2)$
$P(\mathbf{h}^2 = 1|\mathbf{x}^2_{ij}) = \frac{1}{1+\exp(\mathbf{c}+W\mathbf{x}^2_{ij})}$
calculate gradient of Equ. 6 w.r.t $\mathbf{h}$ as $\mathbf{d}_b$
calculate gradient of Equ. 9 w.r.t $\mathbf{h}$ as $\mathbf{d}_s$
$\mathbf{d}_c = (\alpha\mathbf{d}_b + \beta\mathbf{d}_s) \odot P(\mathbf{h}^1 = 1|\mathbf{x}_{ij}) \odot (1 - P(\mathbf{h}^1 = 1|\mathbf{x}_{ij}))$
$\mathbf{D}_w = \mathbf{d}_c\mathbf{x}^T_{ij}$
Update parameters
$W \leftarrow W + \varepsilon(\mathbf{h}^1\mathbf{x}^T_{ij} - P(\mathbf{h}^2 = 1|\mathbf{x}^2_{ij})(\mathbf{x}^2_{ij})^T - \mathbf{D}_w)$
$\mathbf{b} \leftarrow \mathbf{b} + \varepsilon(\mathbf{x}^1_{ij} - \mathbf{x}^2_{ij})$
$\mathbf{c} \leftarrow \mathbf{c} + \varepsilon(\mathbf{h}^1 - P(\mathbf{h}^2 = 1|\mathbf{x}^2_{ij}) - \mathbf{d}_c)$
End While

Output: W, $\mathbf{b}$, $\mathbf{c}$

5 Experiments and Results

5.1 Experimental Setup

We evaluate our algorithm on two real world datasets, PASCAL VOC 2007 dataset (PASCAL for short) [Everingham *et al.*, 2010] and LabelMe dataset [Russell *et al.*, 2008; Liu *et al.*, 2009a]. Extensive comparisons are presented with several related work, including state-of-the-art methods [Xie *et al.*, 2014], [Zhang *et al.*, 2014], [Zhang *et al.*, 2013], [Liu *et al.*, 2012], [Liu *et al.*, 2013], [Vezhnevets *et al.*, 2012], [Vezhnevets *et al.*, 2011], [Vezhnevets and Buhmann, 2010], [Liu *et al.*, 2009b], [Verbeek and Triggs, 2007]. For fair comparison, we directly cite released results of comparison methods. Methods are typically compared using the average per-class accuracy. For a given class, the accuracy is calculated by the percentage of correctly classified pixels.

To further validate the robustness of the proposed method, experiments are performed under the setting of different levels of label noise and different numbers of noise images.

5.2 Results on PASCAL Dataset

The PASCAL VOC 2007 dataset was used for the PASCAL visual object category segmentation contest with 20 object classes [Everingham *et al.*, 2010]. We conduct experiments on the segmentation set with the "train-val" split including 422 training-validation images and 210 test images. For the segmentation dataset, only obvious object labels are provided for each image. As a result, the image-level labels are incomplete to describe the image regions, and large numbers of superpixels are deemed as "background", which leads to diverse backgrounds. It is very challenging to make semantic segmentation on this dataset due to high intra-class variations and diverse backgrounds.

To deal with large numbers of background regions, objectness of each image is calculated to help build semantic graph of object regions [Alexe *et al.*, 2010]. Grabcut [Rother *et al.*, 2004] is adopted to refine the boundary of object regions.

The experimental results of our method compared with other related work are presented in Table 1. The last column shows the average accuracy, and our approach with adaptive block size achieves the best performance with 14.3% relative improvement to state of the art [Xie *et al.*, 2014]. For individual concepts, our approach with adaptive block size and equal block size outperforms the comparison methods on 5 classes and 6 classes respectively. Compared with state of the art [Xie *et al.*, 2014], our approach with adaptive block size is more robust to deal with different kinds of object classes. Although SGC [Xie *et al.*, 2014] achieves best performance in a few classes (e.g., plane, bird and cow), it fails in several classes (e.g., horse, motorbike and dog).

By comparing results of our approach with equal block size and adaptive block size, we can find that the model with adaptive block size is more robust to class changes and achieves better performance by taking data distribution into consideration. The model with equal block size tends to prefer labels with high frequency (e.g. person and chair), while the model with adaptive block size can handle most classes well at the cost of some classes with high label frequency.

5.3 Results on LabelMe LMO Dataset

This dataset is a subset of LabelMe dataset [Russell *et al.*, 2008] provided by [Liu *et al.*, 2009a]. It contains 2688 fully annotated images of 33 object categories, most of which are outdoor scenes including sky, street, buildings,

Table 1: Semantic segmentation results on PASCAL dataset.

Method	plane	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	motorbike	person	plant	sheep	sofa	train	tv	bkgd	mean
[Liu *et al.*, 2009b]	24	25	40	25	32	35	27	45	16	49	24	32	13	25	56	28	17	16	33	18	**82**	32
[Liu *et al.*, 2012]	28	20	52	28	46	41	39	60	25	68	25	35	17	35	56	36	46	17	31	20	65	38
[Zhang *et al.*, 2013]	48	20	26	25	3	7	23	13	38	19	15	39	17	18	25	47	9	**41**	17	33	-	24
[Zhang *et al.*, 2014]	65	25	39	8	17	38	17	26	25	17	**47**	**41**	**44**	32	59	34	36	23	35	31	-	33
[Xie *et al.*, 2014]	**85**	**55**	**87**	45	42	31	34	57	21	81	23	16	6	11	42	31	**72**	24	**49**	40	41	42
Ours (equal block)	56	16	77	25	**62**	22	63	**66**	**42**	**83**	15	37	13	5	**81**	60	50	23	29	**72**	41	45
Ours (adaptive block)	33	50	72	**66**	46	**70**	**73**	43	30	78	29	31	16	**52**	33	**61**	41	38	47	48	50	**48**

Table 2: Semantic segmentation results on LabelMe LMO dataset.

Supervision	Fully supervised setting			
Method	[Shotton *et al.*, 2006]	[Liu *et al.*, 2009a]	[Tighe and Lazebnik, 2010]	[Myeong and Lee, 2013]
Accuracy	13	24	29	32
Supervision	Weakly supervised setting			
Method	[Vezhnevets *et al.*, 2011]	[Vezhnevets *et al.*, 2012]	[Liu *et al.*, 2013]	Ours
Accuracy	14	21	26	**41**

mountain. For the supervised methods [Liu *et al.*, 2009a; Myeong and Lee, 2013; Tighe and Lazebnik, 2010; Shotton *et al.*, 2006], there are 2488 randomly selected images for training and 200 for testing. The occurrence frequency for different labels is highly imbalanced as shown in Figure 2. The label 'sky' occurs in more than 85% images, while the label 'moon' occurs in less than 1% images. Such a dataset is very challenging to make semantic segmentation due to high label imbalance. To deal with the high label imbalance, the block size is adaptively tuned with Equation (7) according to the label frequency.

Comparisons with fully supervised methods and weakly supervised methods are given in Table 2. Our approach achieves even better performance than the fully supervised methods [Myeong and Lee, 2013] by taking label imbalance into consideration. It outperforms [Liu *et al.*, 2013] by 15% improvement and [Myeong and Lee, 2013] by 9% improvement. It can handle most classes well at the cost of some classes with high label frequency like 'sky'. Moreover, For such scene datasets, context information will be helpful to make prediction for the position-fixed classes like 'sky' and 'sea', which will be exploited in future.

Furthermore, to validate the robustness of our model to label noise, we conduct extensive experiments with different numbers of noise labels and noise images. Specifically, we randomly select $p \in \{10, 20, \cdots, 90, 100\}$ percents of images and add $r \in \{1, 2, 3\}$ randomly selected labels. We repeat each experiment 5 times and report the average accuracy. Detailed results can be found in Figure 4. With fixed number of noise label, we find that the average class accuracy decays linearly with the increase of noise images. Our approach achieves 30% accuracy even when every sample is with a noise label. Moreover, the label accuracy drops little with the increase of noise labels (from 1 to 3) when the noise samples are less than 30 percents, since the non-image-level suppression term and semantic graph propagation term help to alleviate the effect of noisy labels. As the number of noise images increases, many cooccurrence of labels may appear, which may have a larger effect on the final result.

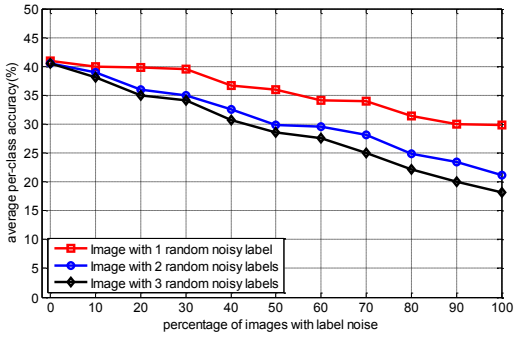

Figure 4: Average per-class accuracy with different numbers of noise labels and noise images on the LabelMe LMO dataset.

6 Conclusion

In this paper, we propose a weakly supervised semantic segmentation method via non-image-level suppression. Hidden units in the WRBM are divided into multiple blocks, and each block corresponds to a specific label. A non-image-level suppression term is imported to suppress the response of blocks with impossible labels, while a semantic graph propagation term is imported to regularize similar features to have similar hidden response. Extensive experiments on two real world challenging datasets demonstrate the good performance of our approach.

7 Acknowledgments

This work was supported by 973 Program (2012CB316304) and National Natural Science Foundation of China (61332016, 61272329 and 61472422).

References

[Achanta *et al.*, 2012] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Ssstrunk. Slic superpixels compared to state-of-the-art superpixel methods. *TPAMI*, 34(11):2274–2282, 2012.

[Alexe *et al.*, 2010] B. Alexe, T. Deselaers, and V. Ferrari. What is an object? In *CVPR*, pages 73–80, 2010.

[Bengio, 2009] Yoshua Bengio. Learning deep architectures for ai. *Foundations and Trends in Machine Learning*, 2(1):1–127, 2009.

[Everingham *et al.*, 2010] Mark Everingham, Luc Van Gool, ChristopherK.I. Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *IJCV*, 88(2):303–338, 2010.

[Farabet *et al.*, 2013] C. Farabet, C. Couprie, L. Najman, and Y. LeCun. Learning hierarchical features for scene labeling. *TPAMI*, 35(8):1915–1929, 2013.

[Hinton and Salakhutdinov, 2006] G. E. Hinton and R. R. Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504–507, 2006.

[Hinton, 2002] Geoffrey E. Hinton. Training products of experts by minimizing contrastive divergence. *Neural Comput.*, 14(8):1771–1800, 2002.

[Ladicky *et al.*, 2009] L. Ladicky, C. Russell, P. Kohli, and P.H.S. Torr. Associative hierarchical crfs for object class image segmentation. In *ICCV*, 2009.

[Ladicky *et al.*, 2010] Lubor Ladicky, Chris Russell, Pushmeet Kohli, and PhilipH.S. Torr. Graph cut based inference with co-occurrence statistics. In *ECCV*, volume 6315, pages 239–253. 2010.

[Liu *et al.*, 2009a] Ce Liu, J. Yuen, and A. Torralba. Nonparametric scene parsing: Label transfer via dense scene alignment. In *CVPR*, 2009.

[Liu *et al.*, 2009b] Xiaobai Liu, Bin Cheng, Shuicheng Yan, Jinhui Tang, Tat Seng Chua, and Hai Jin. Label to region by bi-layer sparsity priors. In *MM*, 2009.

[Liu *et al.*, 2011] Ce Liu, J. Yuen, and A. Torralba. Nonparametric scene parsing via label transfer. *TPAMI*, 33(12):2368–2382, 2011.

[Liu *et al.*, 2012] Si Liu, Shuicheng Yan, Tianzhu Zhang, Changsheng Xu, Jing Liu, and Hanqing Lu. Weakly supervised graph propagation towards collective image parsing. *Multimedia, IEEE Transactions on*, 14(2):361–373, 2012.

[Liu *et al.*, 2013] Yang Liu, Jing Liu, Zechao Li, Jinhui Tang, and Hanqing Lu. Weakly-supervised dual clustering for image semantic segmentation. In *CVPR*, 2013.

[Lowe, 2004] David G. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, 2004.

[Myeong and Lee, 2013] Heesoo Myeong and Kyoung Mu Lee. Tensor-based high-order semantic relation transfer for semantic scene segmentation. In *CVPR*, 2013.

[Rother *et al.*, 2004] Carsten Rother, Vladimir Kolmogorov, and Andrew Blake. "grabcut" - interactive foreground extraction using iterated graph cuts. *ACM Trans. Graph.*, 23(3):309–314, 2004.

[Russell *et al.*, 2008] BryanC. Russell, Antonio Torralba, KevinP. Murphy, and WilliamT. Freeman. Labelme: a database and web-based tool for image annotation. *IJCV*, 77(1), 2008.

[Shotton *et al.*, 2006] Jamie Shotton, John Winn, Carsten Rother, and Antonio Criminisi. Textonboost: Joint appearance, shape and context modeling for multi-class object recognition and segmentation. In *ECCV*, 2006.

[Shotton *et al.*, 2008] J. Shotton, M. Johnson, and R. Cipolla. Semantic texton forests for image categorization and segmentation. In *CVPR*, pages 1–8, 2008.

[Shotton *et al.*, 2009] Jamie Shotton, John Winn, Carsten Rother, and Antonio Criminisi. Textonboost for image understanding: Multi-class object recognition and segmentation by jointly modeling texture, layout, and context. *IJCV*, 81(1):2–23, 2009.

[Tighe and Lazebnik, 2010] Joseph Tighe and Svetlana Lazebnik. Superparsing: Scalable nonparametric image parsing with superpixels. In *ECCV*, 2010.

[Verbeek and Triggs, 2007] J. Verbeek and B. Triggs. Region classification with markov field aspect models. In *CVPR*, 2007.

[Vezhnevets and Buhmann, 2010] A. Vezhnevets and J.M. Buhmann. Towards weakly supervised semantic segmentation by means of multiple instance and multitask learning. In *CVPR*, 2010.

[Vezhnevets *et al.*, 2011] A. Vezhnevets, V. Ferrari, and J.M. Buhmann. Weakly supervised semantic segmentation with a multi-image model. In *ICCV*, 2011.

[Vezhnevets *et al.*, 2012] A. Vezhnevets, V. Ferrari, and J.M. Buhmann. Weakly supervised structured output learning for semantic segmentation. In *CVPR*, 2012.

[Xie *et al.*, 2014] Wenxuan Xie, Yuxin Peng, and Jianguo Xiao. Semantic graph construction for weakly-supervised image parsing. In *AAAI*, 2014.

[Zhang *et al.*, 2012] Dengsheng Zhang, M.Monirul Islam, Guojun Lu, and Ishrat Jahan Sumana. Rotation invariant curvelet features for region based image retrieval. *IJCV*, 98(2):187–201, 2012.

[Zhang *et al.*, 2013] Ke Zhang, Wei Zhang, Yingbin Zheng, and Xiangyang Xue. Sparse reconstruction for weakly supervised semantic segmentation. In *IJCAI*, 2013.

[Zhang *et al.*, 2014] Ke Zhang, Wei Zhang, Sheng Zeng, and Xiangyang Xue. Semantic segmentation using multiple graphs with block-diagonal constraints. In *AAAI*, 2014.